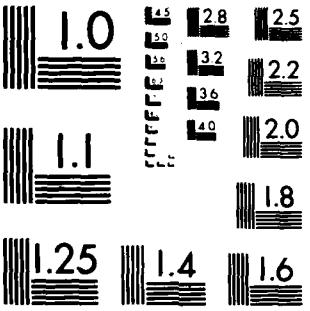
AD-A138 721 A COMPARISON OF SEVERAL ALTERNATIVES TO MAXIMUM LIKELIHOOD FOR THE WEIBUL..(U) UNIVERSITY OF CENTRAL FLORIDA ORLANDO DEPT OF MATHEMATICS AND.. 1/1

UNCLASSIFIED S J BEAN ET AL. 22 SEP 83 AFGL-TR-83-0248 F/G 12/1 NL

END
DATE FILMED
4 84
DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AFGL-TR-83-0248

A Comparison of Several Alternatives to Maximum Likelihood for the Weibull Distribution

Steven J. Bean
Paul N. Somerville
LeRoy A. Franklin

Department of Statistics
University of Central Florida
Orlando, FL 32816

Scientific Report No. 5

22 September 1983

AIR FORCE GEOPHYSICS LABORATORY
AIR FORCE SYSTEMS COMMAND
UNITED STATES AIR FORCE
HANSCOM AFB, MASSACHUSETTS 01731

This report has been reviewed by the ESD Public Affairs Office (PA) and is releasable to the National Technical Information Service (NTIS).

"This technical report has been reviewed and is approved for publication"

FOR THE COMMANDER

DONALD D. GRANTHAM
Chief, Tropospheric Structure Branch

ROBERT A. McCLATCHEY
Director, Atmospheric Sciences Division

Qualified requestors may obtain additional copies from the Defense Technical Information Center. All others should apply to the National Technical Information Service.

If your address has changed, or if you wish to be removed from the mailing list, or if the addressee is no longer employed by your organization, please notify AFGL/DAA, Hanscom AFB, MA 01731. This will assist us in maintaining a current mailing list.

Do not return copies of this report unless contractual obligations or notices on a specific document requires that it be returned.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFGL-TR-83-0248	2. GOVT ACCESSION NO	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) A Comparison of Several Alternatives to Maximum Likelihood for the Weibull Distribution		5. TYPE OF REPORT & PERIOD COVERED Scientific Report No. 5 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Steven J. Bean Paul N. Somerville LeRoy A. Franklin		8. CONTRACT OR GRANT NUMBER(s) F19628-82-K- 0001
9. PERFORMING ORGANIZATION NAME AND ADDRESS University of Central Florida Department of Statistics Orlando, FL 32816		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 62101F 667009AK
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Geophysics Laboratory Hanscom AFB, MA 01730 Contract Monitor: C. F. Burger/LYT		12. REPORT DATE 22 September 1983 13. NUMBER OF PAGES 15
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for Public Release, Distribution Unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Non-Linear Regression
Weighted Least Squares
Generalized Least Squares
Log Linear Model

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Five methods of parameter estimation for the two parameter Weibull distribution are investigated using simulation. They are: Non-Linear Regression, Log-Linearization, Weighted Least Squares, Generalized Least Squares, and Method of Moments. Effects of contamination and censoring of data as well as robustness of the methods were investigated. Weighted Least Squares seemed to be the most cost-effective method and Nonlinear Regression was the most robust.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

A Comparison of Several Alternatives to Maximum Likelihood for the Weibull Distribution

by

Steven J. Bean and Paul N. Somerville
University of Central Florida

1. Introduction

The Weibull distribution is very versatile and has found many uses in reliability, and in the climatological modeling of weather elements. In this study, we compare several alternatives to maximum likelihood (ML) estimation. Somerville and Bean (1982) compared ML and least squares (LS) and found that under ideal conditions LS and ML gave substantially the same results. However, when contamination or censoring occurs, or when the wrong model is used, LS can give substantially better results. In addition, ML estimation of the parameters of the Weibull distribution requires iterative techniques. The alternatives in this study are considered for two reasons. They are more robust, and most of them are much easier to compute.

It is common to evaluate estimators on the basis of their variances and biases. Although these are important considerations, a user is frequently more interested in how well the model is going to predict probabilities. That is, we are more interested in the fit of the cumulative distribution than the values of the parameters. We will evaluate the alternatives on the basis of the fit to the cumulative distribution. We use the following form of the cumulative distribution function

$$F(x) = 1-\exp(-\alpha x^{\beta}), \qquad x,\alpha,\beta > 0. \tag{1}$$

2. Alternatives to Maximum Likelihood

Let $x_1, x_2, \ldots, x_n$ be the ordered observations of a random sample from the Weibull distribution. We use the following form for the empirical cumulative distribution function (CDF)

$$F_n(x_i) = (i-.5)/n \tag{2}$$

and

$$F_n(x) = F_n(x_i) \qquad x_i \leq x < x_{i+1}. \tag{3}$$

The estimators which follow (except for the method of moments) select (α,β) so as to minimize the "distance" between $F(x;\alpha,\beta)$ and $F_n(x)$.

2.1 Non-Linear Regression (NLR)

In non-linear regression, (α,β) are selected so that the expression

$$\sum_{i=1}^{n} (F(x_i;\alpha,\beta) - F_n(x_i))^2 \tag{4}$$

is minimized. This ensures that the model distribution fits the empirical CDF in the least squares sense. However, costly iterative techniques are required.

2.2 The Log-Linearization Method (LLM)

The non-linear model in 2.1 may be made linear by using logarithms. Let

$$q_i = 1 - F_n(x_i) \tag{5}$$

$$\hat{q}_i = \exp(-\hat{\alpha} x_i^{\hat{\beta}}) \tag{6}$$

where $\hat{\alpha}$ and $\hat{\beta}$ are estimates of α and β. Then

$$\ln(-\ln \hat{q}_i) = \ln \hat{\alpha} + \hat{\beta} \ln x_i. \tag{7}$$

We may regard this as a regression model with $\ln(-\ln q_i)$ as the dependent variable and $\ln x$ as the independent variable. Reversing the roles of dependent and independent variables, we may also write

$$\ln \hat{x}_i = \ln(-\ln q_i)/\hat{\beta} - (\ln \hat{\alpha})/\hat{\beta}. \tag{8}$$

Ordinary least squares may be used to obtain model coefficients from which estimates of (α,β) may be obtained. Using (7), the model attempts to fit the CDF while in (8) the model attempts to fit the percentiles. However, using equations (5) and (6), the non-linear least squares equation (4) can be written as

$$\sum_{i=1}^{n} (q_i - \hat{q}_i)^2 \tag{9}$$

and it is this sum (9) that non-linear least squares seeks to minimize. Using the log-linearization method coupled with ordinary least squares one is seeking to minimize

$$\sum_{i=1}^{n} (\ln(-\ln q_i) - \ln(\ln \hat{q}_i))^2 \qquad (10)$$

Since the sum of squares being minimized is different, the resulting estimates of $\hat{\alpha}$ and $\hat{\beta}$ derived from log-linearization can be very different from the estimates derived using non-linear regression.

2.3 Weighted Least Squares (WLS)

If we can weight the regression using the log-linearization method in such a way that the weighted distance metric is the same as for the non-linear regression method, then we can achieve the results of the non-linear regression without using the costly iterative technique. That is, by putting $u = \ln(-\ln q_i)$, $v = \ln x_i$, $a = \ln \alpha$ and $b = \beta$, (7) becomes $u = a + bv + e$ and we solve for values of a and b which minimize $\Sigma w_i^2 e_i^2$.

Also, a weighted log-linear result using (7) and using weights w_i may be obtained by minimizing

$$\sum_{i=1}^{n} w_i^2 (\ln(-\ln q_i) - \ln(\ln \hat{q}_i))^2 \qquad (11)$$

Equating expressions (9) and (11), and solving for w_i we have

$$1/w_i = (\ln(-\ln q_i) - \ln(-\ln \hat{q}_i))/(q_i - \hat{q}_i) \qquad (12)$$

As $$\hat{q}_i \to q_i,\ 1/w_i \to (\ln(-\ln q_i))' \qquad (13)$$

where the prime indicates the derivative with respect to q_i, and we have $w_i = -q_i \ln q_i$.

Thus, using (11) with $w_i = -q_i \ln q_i$, we have a weighted least squares method (using ordinary least squares) which is approximately equivalent to using *non-linear* least squares. (Similar results are obtained when the roles of the independent and *dependent* variables are reversed and (8) is used. In that case it is easy to show that $w_i = x_i$.)

2.4 Generalized Least Squares (GLS)

Engeman and Keefe (1982) describe a method which takes into account the fact that the observations must be ordered to calculate the empirical CDF or percentiles. Putting

$$\hat{y}_i = \ln \hat{x}_i,\ \hat{z}_i = \ln(-\ln q_i),\ \hat{a} = -(\ln \hat{\alpha})/\hat{\beta}, \hat{b} = 1/\hat{\beta} \qquad (14)$$

(8) may be written as

$$\hat{y}_i = \hat{a} + \hat{b}\, z_i. \qquad (15)$$

However, the variance-covariance matrix of $Y' = (y_1, y_2, \ldots, y_n)$ is not $\sigma^2 I$ but $\sigma^2 V$. An approximation to V is given by

$$v_{ij} = (1-q_i)/(q_i \ln q_i \ln q_j) \quad \text{for } i \le j. \qquad (16)$$

The resulting estimates from Engeman and Keefe are obtained from the equation

$$(\hat{a},\hat{b})' = (Z'V^{-1}Z)^{-1}(Z'V^{-1}Y), \tag{17}$$

where Z consists of a column of ones and a column of the values $\ln(-\ln q_i)$. Knowing estimates of a and b, we can find estimates of α and β since

$$\begin{aligned} \hat{\alpha} &= \exp(-\hat{a}/\hat{b}) \\ \hat{\beta} &= 1/\hat{b}. \end{aligned} \tag{18}$$

2.5 Methods of Moments (MM)

Menon (1963) used the method of moments to obtain

$$\begin{aligned} \hat{b} &= (6s^2/\pi^2)^{1/2} \\ \hat{a} &= \bar{x}' + /5772\ b \end{aligned} \tag{19}$$

where s^2 and $\bar{x}'$ are the sample variance and mean, respectively, of $\ln x_i$. We then calculate

$$\begin{aligned} \hat{\alpha} &= \exp(-a/b) \\ \hat{\beta} &= 1/b. \end{aligned} \tag{20}$$

3. Simulation Results

The five estimation methods described in the previous section, the log-linear (LL), weighted least squares (WLS), non-linear regression (NLR), generalized least squares (GLS), and Menon's methods of moments (MM) were used on a number of generated data sets. All of the methods except MM could be approached from two directions. The ln x could be considered the dependent variable or ln(-ln q) could be considered as the dependent

variable. The LL, WLS, and NLR generally gave better results when $\ln(-\ln q)$ was used as the dependent variable. GLS gave much better results when $\ln x$ was used as the dependent variable. Therefore, we present results for $\ln(-\ln q)$ as the dependent variable when LL, WLS, or NLR was used, and $\ln x$ when GLS was used.

Samples of size $n = 25$ were generated from each of the distributions in the study. For each distribution $N = 30$ replications were used. Several Weibull distributions were used with $\alpha = .001, .01, .1, 1, 10$ and $\beta = .5, 1, 2, 4$. Since a number of the combinations yielded nearly the same results, we report only the following (α,β) combinations: (1,1), (10,1), (1,2), and (1,4).

Contamination and censoring are two factors which can significantly affect the estimates. One form of contamination was simulated by randomly taking 8% or 2 out of 25 of the observations and transforming them by $a + \sqrt{b}x$. Values of (a,b) were (1,1), (.5,2) and (0,4). Only $a = .5$, $b = 2$ is reported in Table 3.2 since the other values yielded similar results and this set of values gives the most contrast between the methods. Censoring was simulated by keeping a percent of the smallest observations. We used 40%, 60%, and 80% as percentages of observations kept.

The robustness of the estimation methods was also tested by using the Weibull model on data generated from another distribution. We used the normal distribution with $\mu = 3.5$, $\sigma^2 = 1$ and $\mu = 6$, $\sigma^2 = 1$, the log-normal with $\mu = 0$, $\sigma^2 = 1$, and the "log-Cauchy" distribution. The log-Cauchy distribution was obtained by generating a Cauchy observation x and transforming it by e^x. The location and scale parameters were 0 and .1, respectively.

Error evaluations were recorded for all of the different cases. The root mean square (RMS) was calculated with respect to the true underlying distribution, and the RMS of the parameters was also calculated. The formulas used are given by

$$RMS = \left(\sum_{j=1}^{N} \sum_{i=1}^{n} (F(x_{ij}) - F(x_{ij};\alpha_j,\beta_j))^2/Nn \right)^{1/2} \tag{21}$$

$$RMS_\alpha = \left(\sum_{j=1}^{N} (\alpha - \alpha_j)^2/N \right)^{1/2} \tag{22}$$

$$RMS_\beta = \left(\sum_{j=1}^{N} (\beta - \beta_j)^2/N \right)^{1/2}. \tag{23}$$

$F(x_{ij})$ is the value of the true underlying distribution for the i^{th} observation in the j^{th} generated sample. The sample estimates for α_j and β_j are α and β using the j^{th} sample. Tables 3.1 through 3.3 give the results for the simulated problems.

TABLE 3.1

RMS, RMS_α, RMS_β For the Fit to the Known Weibull Distribution
(All combinations of (α,β) gave essentially the same results)

	LL	WLS	GLS	NLR	MM
RMS	.056	.061	.055	.064	.059
RMS_α	.230	.228	.221	.263	.252
RMS_β	.216	.212	.194	.251	.252

TABLE 3.2

RMS, RMS_α, RMS_β For the Fit to the Known Weibull Distribution When 8% of the Generated Values Have Been Replaced by a Contaminated Value

	LL	WLS	GLS	NLR	MM
			($\alpha = 1, \beta = 1$)		
RMS	.060	.059	.057	.060	.060
RMS_α	.202	.183	.199	.194	.210
RMS_β	.206	.200	.194	.277	.233
			($\alpha = 10, \beta = 1$)		
RMS	.074	.065	.072	.063	.072
RMS_α	3.81	4.01	4.39	5.75	3.68
RMS_β	.150	.160	.179	.197	.147
			($\alpha = 1, \beta = 2$)		
RMS	.072	.062	.069	.061	.070
RMS_α	.228	.180	.208	.187	.228
RMS_β	.348	.369	.369	.402	.359
			($\alpha = 1, \beta = 4$)		
RMS	.122	.080	.116	.063	.112
RMS_α	.328	.186	.269	.192	.324
RMS_β	1.16	1.15	1.50	.78	.87

Note: A contaminated observation is .5 + 2x, where x is the true observation: all other contaminations gave similar results.

TABLE 3.3

RMS Values For the Fit to the Known Weibull Distribution Where a Percentage of the Largest Values Have Been Removed or Censored

Distrib.	% Censored	LL	WLS	GLS	NLR	MM
	60	.123	.107	.100	.105	-
Weibull	40	.089	.080	.079	.081	-
α=1,β=4	20	.070	.067	.063	.069	-
	0	.056	.061	.055	.064	.059
	60	.104	.092	.097	.091	-
Normal	40	.080	.074	.076	.075	-
N(6,1)	20	.064	.066	.064	.065	-
	0	.061	.062	.059	.063	.063
	60	.120	.099	.111	.113	-
Log Normal	40	.091	.078	.082	.080	-
μ=0,σ^2=1	20	.072	.071	.070	.069	-
	0	.071	.067	.067	.067	.073
Log Cauchy*	0	.165	.118	.132	.075	.153

*The Log-Cauchy was formed using the Cauchy probability density function for x given by $s/(x^2 + (x-m))^2$ where x = .1, m = 0. If x has a Cauchy pdf, then the Log-Cauchy variable is exp(x).

Table 3.1 gives a comparison of the various methods under ideal conditions. That is, the model distribution is correct, and there is no contamination or censoring of the data. The values are essentially the same for all (α,β) combinations with the exception of α = 10, β = 1, in

which case the RMS_α is approximately 20 times larger than other RMS, RMS_α, and RMS_β values but the pattern is the same. The GLS method out-performs the other methods with respect to the fit of the CDF and the estimation of (α,β). It is interesting to note that the LL method is close to GLS. This is important for cases in which computer storage space is limited and the sample size is large since GLS does require an additional nxn matrix.

Table 3.2 illustrates the effects of contamination. WLS and NLR appear to be very stable with respect to the fit of the CDF. The LL, GLS, and MM tend to degrade for larger α or β values.

Table 3.3 shows that GLS does slightly better than WLS and NLR when the underlying distribution is Weibull and censoring is present. Also worth noting is that the LL method appears to be more adversely affected by censoring than the other methods.

Table 3.3 also shows the robustness properties of the estimators. It is interesting that the Weibull can approximate a normal distribution quite well when no censoring is present for any estimation methods. Another interesting case is that of the log-Cauchy distribution, Only NLR was able to give satisfactory results, and actually the fit is surprisingly good considering the underlying distribution.

4. SUMMARY AND CONCLUSIONS

We have examined the small sample properties of five methods of parameter estimation for the two parameter Weibull distribution using simulation. It is clear that the GLS is the best estimator and LL is a close second best with respect to the fit of the CDF and the RMS of the parameters under ideal conditions. When contamination is present, WLS and NLR are very stable whereas the other methods give much poorer results

especially when α or β are larger than 2. NLR also gives much better results than the other methods when the underlying distribution is the log-Cauchy distribution. WLS, GLS, and NLR perform well when the underlying distribution is the Log-Normal distribution.

Computationally LL and MM are certainly the least expensive. They are easy to use even on a hand calculator, and the results they give are not far from the GLS under ideal conditions. However, LL and MM do not offer the robustness of the other three methods. The robustness of the WLS, GLS, and NLR costs in terms of complexity and storage. NLR requires a fairly complex program or a software package to implement. GLS requires an additional nxn matrix which can be prohibitive for large sample sizes. This problem can probably be overcome bv grouping data if storage is a problem. WLS requires only an additional n weights to the LL method.

The WLS method appears to be the most cost effective method of the five examined in this study. The NLR method appears to be the most robust. If we consider the log-Cauchy to be an extreme case of a wrong model using the Weibull, then we may conclude that NLR is an extremely robust procedure.

5. ACKNOWLEDGEMENTS

The research was motivated and partially supported by Contract No. F19628-82-K-0001 with the Air Force Geophysics Laboratory.

The method which we denote by WLS (weighted least squares) was first suggested to us by Major A. Boehm of the U.S. Air Force.

All the programs used to make the simulations reported on in this manuscript were written by David Van Brackle. The programs were written in FORTRAN 77 and run on a DEC-VAX-11/780.

6. REFERENCES

Engeman, R. M. and T. J. Keefe 1982: Generalized Least Squares Estimation of the Weibull Distribution, Communications in Statistics, Theory and Methods, 11(19), 2181-2193.

Menon, M. H. 1963: Estimation of the Shape and Scale Parameters of the Weibull Distribution, Technometrics 5, 175-182.

Somerville, P. N. and S. J. Bean 1982: A Comparison of Maximum Likelihood and Least Squares For the Estimation of a Cumulative Distribution, J. Statist. Comput. Simul. Vol. 14, 229-239.